DOCUMENT RESUME

ED 333 032 TM 016 535

AUTHOR Myford, Carol M.
TITLE Judging Acting Ability: The Transition from Novice to Expert.
PUB DATE Apr 91
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991). For a related document, see TM 016 536.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *Acting; *Aesthetic Values; Art Criticism; Comparative Analysis; Drama; Evaluation Methods; *Evaluators; High Schools; High School Students; Individual Development; Interrater Reliability; Matched Groups; Secondary School Teachers; Student Evaluation; *Value Judgment; Videotape Recordings
IDENTIFIERS *Experts; Performance Based Evaluation

ABSTRACT

The aesthetic judgments of experts (casting directors and high school drama teachers), theater buffs, and novices were compared as they rated high school students' videotaped performances of Shakespearean monologues. It was hypothesized that theater buffs would represent an intermediate stage on the path to developing expertise in judging acting ability. The judge sample (N=27) included nine experts, nine theatre buffs, and nine novices, with each expert being matched with a theatre buff and novice of the same sex and approximately the same age and level of education. All of the judges viewed eight high school students' videotaped performances of 2-minute long monologues twice, rated the videotapes, and completed the 36-item Judging Acting Ability Inventory developed for this study. One month later, each judge viewed the same eight videotapes of the student performances twice, and again completed the rating and sorting tasks. Theater buffs did represent an intermediate stage in the development of expertise in judging acting. Their measures of contestant ability were significantly different than those of the experts and novices, with more similarities to the ratings of experts. Theatre buffs were also better at replicating their results at a second session than were novices, but they did not perform as well as did the experts. Implications of the results for the judgment of aesthetic experience are discussed. Three tables, two graphs, and a 38-item list of references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CAROL M. MYFORD

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED333032

Judging acting ability: The transition from novice to expert

Carol M. Myford
Educational Testing Service, Princeton, NJ

Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April, 1991

TM016535

BEST COPY AVAILABLE

Judging acting ability: The transition from novice to expert

Does expertise in making aesthetic judgments exist? If so, what is the nature of expertise in performing this task? How do the aesthetic judgments of experts differ from those of novices? In the past, researchers investigating these questions have used the term "expert" freely, employing a variety of operational definitions for the construct. While they seem to agree that an expert has specialized training and experience in a field, they disagree over the issue of just how much specialized training and experience one must have in order to qualify as an expert. Consequently, persons who qualify as experts in one study would clearly not meet the selection criteria for another study.

A high school student studying art served as one of the experts in Beard's (1978) study, while undergraduate or graduate students majoring in art are the experts in other studies (Bamossy, Johnston & Parsons, 1985; Child, 1962; Eysenck, 1972). Still other studies have used persons who work in art-related professions, such as art teachers, practicing artists, art historians, art critics, and museum directors (Burt, 1934; Cattell, Glascock & Washburn, 1918; Getzels & Csikszentmihalyi, 1969; Gordon, 1952; Skager, Schultz & Klein, 1966; Wilson, 1970). Curiously, researchers make little or no mention of the extent of an expert's experience in the field or the amount of specialized training the expert has had. There seems to be little acknowledgment that persons with more training or experience might be more expert than persons with less training or experience, or that differing patterns of education and experience in art might produce qualitatively different kinds of art-related expertise (i.e., an art historian's expertise might differ from that of a practicing artist).

There is also a great deal of controversy in the aesthetic judgment literature over the definition of the construct "novice." Researchers use several different terms to describe novices: "non-experts" (Getzels &

Csikszentmihalyi, 1969), "laymen" or "lay judges" (Eysenck, 1972; Gordon, 1952), "non-artists" (O'Hare, 1976), "non-sophisticated subjects" (Berlyne & Ogilvie, 1974), and "the untrained" (Burt, 1934). The criteria researchers use for selecting persons to serve as novices differ markedly. In some studies elementary, junior high, or senior high school students serve as novices (Bamossy, Johnston & Parsons, 1985; Burt, 1934; Dewar, 1938), while in other studies undergraduate or graduate students majoring in subjects other than art are used as novices (Berlyne & Ogilvie, 1974; Child, 1962; Eysenck, 1972; Getzels & Csikszentmihalyi, 1969). Other researchers have employed college faculty members who teach non-art subjects (Beard, 1978), factory workers (Frances & Voillaume, 1964; Hussain, 1966; Voillaume, 1965), or other adults not engaged in arts-related professions (Burt, 1934; Skager, Schultz & Klein, 1966).

In the past researchers have most often compared experts' and novices' ratings (or rankings) of works of art in order to gain insight into the nature of expertise. Only a few researchers (Bamossy, Johnston & Parsons, 1985; Burt, 1934; Cattell, Glascock & Washburn, 1918; Frances & Voillaume, 1964; Voillaume, 1965) have attempted to examine the ratings (or rank orderings) made by groups at various points along the continuum--persons who are clearly not novices, since they have some specialized training and experience in art; but who are also clearly not experts, since they lack the depth and breadth of training and experience in art that experts would possess.

If we are to understand how expertise develops so that we can explain the mechanisms by which one makes the transition from novice to expert status, then we need to understand what characterizes performance at various points along the expertise continuum. We must adopt a developmentalist's perspective and pose new questions to guide our research: Are there stages in the development of expertise in making aesthetic judgments? If so, can we

define those stages? How do we identify persons who are in the various stages? What characterizes their rating behavior (i.e., how does it differ from that of experts and of novices)? What triggers movement from one stage to the next?

The purpose of the present study was to compare the aesthetic judgments of experts (i.e., casting directors and high school drama teachers), theater buffs, and novices as they rated high school students' videotaped performances of Shakespearean monologues. The investigator hypothesized that theater buffs would represent an intermediate stage on the path to developing expertise in performing this task. The three judge groups' ratings of the performances were compared to determine whether the theater buffs did indeed function as a transitional group. The study's focus upon gaining an understanding of the nature of expertise in evaluating acting ability extends the scope of aesthetic judgment research beyond the visual arts to drama.

A goal of the study was to identify objective criteria that constitute "some necessary, if not sufficient, conditions for defining expertise within a given situation" (Einhorn, 1974, p. 562). In the past visual arts researchers have investigated only a limited number of criteria that might differentiate the ratings of experts and novices. Some have hypothesized that experts would show stronger agreement in their aesthetic responses to works of art than novices. Valentine (1962), Child (1968, 1972) and Winner (1982) have reviewed the literature on inter-judge reliability as a criterion for expertise. The results are mixed, with some studies showing strong between-judge agreement in the ratings of experts while other studies reveal a lack of agreement between experts' ratings.

A few researchers have suggested that experts should be better able to replicate their ratings than novices (Bamossy, Johnston & Parsons, 1985;

Beard, 1978; Dewar, 1938; Einhorn & Koelb, 1982; Farnsworth, 1969; Gordon, 1923; Skager, Schultz & Klein, 1966). With the exception of Beard, the researchers presented test-retest reliabilities for experts but not for novices. The researchers found that experts reproduced their ratings with a high degree of accuracy (i.e., correlations in the range of 0.7 to 0.9 depending upon the study). However, since researchers did not present test-retest reliabilities for novices, there was no basis for comparison to determine whether novices could reproduce their ratings with the same degree of accuracy. Beard (1978) did gather data to allow such a comparison. When Beard compared the two sets of rating data, he found that experts had higher test-retest reliabilities than novices.

Myford (1989) proposed a number of criteria that might differentiate the three groups' ratings of the students (i.e., contestants). Nine criteria were tested. The present study reports on the results obtained for two of those criteria. (See Myford (1989) for a description of the other seven criteria and the results obtained for each.)

Criterion 1. Are the contestant ratings for experts, buffs, and novices significantly different? When judges rate contestants, they will give some performances higher marks than others. The contestants can be ordered by ability from lowest to highest. Perhaps the three groups' orderings of contestants' performances differ. The groups may not define good acting in the same manner. What one group considers good acting another group might consider poor acting.

Criterion 2. Do experts, buffs, and novices differ in their ability to replicate their ratings of contestants' performances? In this study, the judges rated the same set of contestants on two occasions one month apart. The investigator hypothesized that experts, buffs, and novices may differ in the

stability of their ratings. Experts may show stronger evidence of ability to replicate their ratings than buffs, but buffs may be better able to replicate their ratings than novices.

Method

Subjects

The judge sample (N = 27) was composed of nine experts, nine theater buffs, and nine novices. Two groups of experts were included in the study: casting directors and high school drama teachers. Four of the casting directors had each cast at least four Equity theater productions in Chicago, and two had cast for film and television. The four drama teachers each had at least 18 years of experience teaching high school drama. The theater buffs attended the theater regularly, seeing on average 10 performances a year. None of the novices were frequent theatergoers. They typically attended about one live performance a year.[1]

A matched subjects design was employed. Since the subjects were not randomly selected, matching was used to control for the effects of age, sex, and educational level across the three groups. Each expert was matched with a buff and novice of the same sex and approximately the same age and level of education. The average age of the novices was 42.11 years (SD = 5.87), while the buffs' average age was 41.11 years (SD = 9.34). The experts' average age was 42.56 years (SD = 4.77).

Experts in this study were casting directors and high school drama teachers practiced in their craft who had logged many hours in evaluating actors' abilities. Each had formal training in drama and was fluent in the

---

[1]Each judge completed a questionnaire to provide information about his/her background. See Myford (1989) for comparisons of the experts', buffs', and novices' formal training in drama and their drama-related experiences.

language of the discipline. The experts were very familiar with the criteria used in judging acting ability and made such judgments routinely as part of their job assignments. They had experience working with actors of various ages and abilities including teenage actors.

Theater buffs who participated in the study were not formally trained in the discipline but attended professional theater regularly, read reviews, enjoyed talking about drama, and had some knowledge of the kinds of criteria used in evaluating acting. While they may have spent time discussing with others the merits and shortcomings of actors they had seen, they had neither the breadth nor depth of experience in critically analyzing performances that the experts had. Furthermore, while all the buffs attended professional productions, they infrequently viewed high school productions. It was hypothesized that the buffs represented an intermediate stage in the development of expertise in judging acting ability.

Novices in this study were persons who attended the theater very infrequently, rarely read critics' reviews of theatrical performances, and had little training or experience in drama beyond high school. They lacked knowledge of the technical vocabulary used in talking about acting and had no formal experience judging actors' abilities.

Materials

Videotapes

The judges rated eight high school students' videotaped performances of monologues from Shakespearean tragedies and history plays. Each monologue lasted approximately two minutes. All contestants' videotapes conformed to certain standards in order to control for extraneous differences between them (e.g., no character costumes, makeup, or changes in lighting, etc.). All

contestants were taped against a neutral backdrop using one fixed camera at a fixed angle with a fixed lens.

The eight monologues were copied on to four master tapes. All tapes contained the same monologues, but the order of the monologues differed across tapes to counterbalance the presentation of the monologues across judges.

Judging Acting Ability Inventory

The judges rated monologue performances using the Judging Acting Ability Inventory which consists of 36 items, each item describing a standard of good acting. The investigator designed the rating instrument in collaboration with casting directors and drama teachers. Eleven items are designed to assess the actor's voice. Eleven items assess the actor's body, and fourteen items assess the actor's characterization. Judges determine whether the student performs well or poorly on each standard and then decide how well or how poorly. All items use a common six-point rating scale with the points defined as "very poorly," "moderately poorly," "slightly poorly," "slightly well," "moderately well," and "very well." Judges circle their response to each item.

Procedure

Each judge met individually with the investigator for an hour. The judges viewed the performances twice--once to become familiar with the actor and the monologue, and the second time to rate each performance. Two tapes were used to counterbalance the presentation of monologues across judges. The investigator stopped the videotape after presenting each monologue to allow the judge to fill out the Judging Acting Ability Inventory for the contestant. After rating the eight performances, the judge sorted them into

categories and then ordered the performances within each category from best to worst.

Each judge returned for a second rating session one month later to gather data to examine the question of replicability. Again, each judge saw the eight performances twice: the first time to become re-acquainted with the monologues, and the second time to rate the performances. The tapes were counterbalanced in the second session as in the first. The judge then sorted the performances into categories. After completing the rating and sorting tasks, the judges filled out a short questionnaire describing their education and experience in drama.

Results

The rating data were analyzed using a Rasch rating scale computer program called FACETS (Linacre, 1989). FACETS was developed to handle the complexities of many-faceted data. In this study the data have four facets: (1) rating items, (2) contestants, (3) judges, and (4) rating occasions. Information about each of these facets is needed in order to understand the subtleties of the rating situation. FACETS provides a means of investigating each of these facets independent of the other facets.

Several FACETS analyses were run on the rating data. Separate FACETS analyses were run for each group for each rating session (e.g., an analysis of the experts' ratings of contestants for Time 1 and a separate analysis of the experts' ratings of contestants for Time 2).

Criterion 1. Are the contestant ratings for experts, buffs, and novices significantly different? The FACETS program produces an estimate of each

contestant's ability in logit units[2] called a contestant "measure" which is computed from the judges' ratings of the contestant. The higher the contestant measure, the greater the contestant's ability. Contestant measures were computed separately for each group of judges, and the three sets of contestant measures were compared to ascertain whether there were performances which the groups rated differently.

Omnibus chi-square tests of rating consistency[3] were run to determine whether the contestant measures varied significantly across groups at Time 1 and at Time 2. The contestant measures for the three judge groups were significantly different both at Time 1 ($\chi_{16}^2 = 593.12$, $p < .001$) and at Time 2 ($\chi_{16}^2 = 599.46$, $p < .001$).

Pairwise tests for rating consistency were run to find out where the between-group differences lay. The results displayed in Table 1 show that each groups' contestant measures were significantly different from the other two groups' contestant measures for both rating occasions. The largest difference was between experts' and novices' contestant measures, while the

---

[2]Logit units are used rather than raw score units because raw scores are nonlinear ordinal-level data. Arithmetical operations can't be performed on ordinal-level data, since the operations assume equal intervals. Consequently, raw scores must undergo a linear transformation to convert them to an equal-interval metric, such as the logit.

[3]The chi-square test for rating consistency is an analogue to Hedges & Olkin's (1985, p. 123) test for homogeneity of effect sizes. Chi-square tests for rating consistency were used rather than traditional analysis of variance methods to test for significant differences in the three groups' contestant measures. Each contestant measure has a standard error associated with it, and the computation of the chi-square statistic takes into consideration each measure's standard error. By contrast, analysis of variance techniques assume that the error variance for the contestants is distributed identically and independently over all the measures, not acknowledging that individual contestant measures may have different standard errors. Because the chi-square test for rating consistency makes use of more information about each contestant (i.e., both the contestant measure and the standard error for the measure), this methodology was selected over traditional analysis of variance techniques. For details of the technique, see Myford (1989).

smallest difference was between experts' and buffs' measures. Buffs' measures of contestant ability were more like the experts' measures than the novices' measures at both Time 1 and Time 2.

Insert Table 1 about here

Which contestants did the groups rate differently? A chi-square test for rating consistency was run for each individual contestant to pinpoint those particular contestants whom the three judge groups viewed differently. Tables 2 and 3 present the results of those analyses. The chi-square values have been converted into *z* scores by taking the square root of each chi-square value. (The same information is presented in Figure 1 but in a pictorial format that more clearly displays the continuum of contestant ability. In Figure 1 each contestant measure is bracketed by its standard error.) The three groups differed in the estimations of various contestants' abilities as shown in Tables 2 and 3. There were some contestants that buffs rated more like the novices did (i.e., Caliban and Iago at Time 1 and Caliban, Iago, and Mercutio at Time 2) and certain other performances that the buffs rated more like the experts did (i.e., Mark Antony at Time 1 and Mark Antony and Juliet at Time 2). By contrast, experts' and novices' ratings of all contestants at Time 1 and all contestants except Iago at Time 2 were significantly different.

Insert Tables 2 and 3 about here

How did the groups' contestant measures differ? Did they order contestants by ability differently? To the contrary, Figure 1 shows that the contestant orderings for the three groups were similar. Each group's ordering

shows a progression from Mercutio and Paulina at the lower end of the acting ability continuum to Caliban, Ophelia, and Mark Antony at the upper end of the continuum. Only in the case of the Lady Anne portrayal was there a decided difference of opinion about the placement of this performance in comparison to the others. Buffs and experts gave the Lady Anne portrayal significantly lower ratings than novices did. With the exception of the Lady Anne performance, then, the groups seem to share a common definition of what constitutes "good" and "poor" acting.

Where the groups seem to differ is in their judgments of just how good or how poor a performance is. This is particularly noticeable in the cases of the Lady Anne, Mark Antony, and Ophelia performances. For these three contestants the novices' ratings were markedly higher than the buffs' and experts' ratings.

Insert Figure 1 about here

Criterion 2. Do experts, buffs, and novices differ in their ability to replicate their ratings of contestants' performances? Contestant measures for each judge group for each rating occasion were compared to determine whether the groups differed in their abilities to replicate their ratings of contestants' performances. An omnibus chi-square test of the consistency of contestant ratings produced a χ_{24}^2 value of 217.95 which is significant at the .001 level. The three judge groups differed in their abilities to replicate their ratings of contestants from Time 1 to Time 2.

How much change occurred from Time 1 to Time 2 for each group? Chi-square tests for each group across the two rating occasions revealed that all three groups showed significant change in their ratings from one session to

the next, but the amount of change differed from group to group. Novices' ratings changed the most ($\chi_8^2 = 122.73$, $\rho < .001$). By contrast, experts' ratings changed the least ($\chi_8^2 = 32.06$, $\rho < .001$). Buffs' ratings also changed significantly from Time 1 to Time 2 ($\chi_8^2 = 63.166$, $\rho < .001$). The amount of change for buffs was nearly twice that for experts, while the amount of change for novices was nearly four times that for experts.

Which contestant measures changed the most from Time 1 to Time 2? Figure 2 compares the changes in measures of contestant ability from Time 1 to Time 2 for the three groups. For each contestant, the standardized difference of each contestant's measure across the two occasions for experts, buffs, and novices is shown. Points outside the range of +2 to -2 standard errors denote significant change in that group's rating of the contestant across times. All contestants except Mark Antony show significant change for at least one judge group from Time 1 to Time 2.

Insert Figure 2 about here

The greatest amount of change across times was in the novices' and buffs' ratings of the Ophelia performance. There were about 8-1/2 standard errors difference between the novices' contestant measures for Ophelia at Time 1 and at Time 2, while there were nearly 6-1/2 standard errors difference between the buffs' contestant measures of the same performance. The experts also rated the performance significantly higher the second time, but the amount of change was not as great (i.e., about 2-1/2 standard errors difference between experts' contestant measures for Ophelia from Time 1 to Time 2). There was also much change in the novices' ratings of the Juliet performance

from Time 1 to Time 2 (i.e., about 6 standard errors difference) but much less change over time for experts' and buffs' ratings of Juliet.

Discussion

The study produced several sources of evidence that suggest that theater buffs represent an intermediate stage on the path to developing expertise in judging acting ability. First, buffs' measures of contestant ability were significantly different from both experts' and novices' measures of contestant ability. Overall, buffs' ratings were more like experts' ratings than novices', but buffs' ratings of certain performances were more like the novices' ratings than the experts' ratings. Second, buffs were better at replicating their ratings than novices, but they were not as good as experts. These two sources of evidence suggest that the theater buff is indeed in transition--no longer a novice, but not yet an expert.

Why would buffs rate only certain performances like experts do and not others? Why are experts better at replicating their ratings than buffs? What is it that the expert sees in a performance that the buff does not yet see? What has triggered the movement from novice to buff status, and what will it take to move the buff further along the expertise continuum? In order to answer these questions, researchers will need to design studies which probe the cognitive processes novices, buffs, and experts engage in as they rate a performance. Landy and Farr (1983) reviewed the performance rating literature in industrial and organizational psychology and concluded that we know comparatively little about the actual process of making judgments, since no researchers prior to 1983 had investigated raters' cognitive processes. The vast majority of performance rating studies examined the products of those judgments--ratings or rankings--not the judgment process. Had Landy and

Farr reviewed the aesthetic judgment literature, they would have reached much the same conclusion. While visual arts researchers have studied experts' and novices' ratings and rankings of works of art since the early 1900's, few have examined the cognitive processes judges engage in as they produce those judgments. If we are to understand how one moves from novice along the continuum to expert, then we will need to examine the kinds of changes which take place in the judge's ability to process, categorize, store, and recall information.

It may be difficult to design such research in the absence of a theoretical framework that describes how the expert makes judgments about the quality of acting ability. Therefore, a theoretical view of the nature of expertise in performing this task will be proposed. The following discussion is based on the issues raised in this study and is not intended to provide a complete description of how this form of expertise operates. Rather, it represents a first attempt to set out a formal theory which can subsequently be modified and extended as researchers gain insights into the cognitive processes judges employ as they evaluate acting ability.

The casting directors and drama teachers in this study had extensive formal training in their field. Many had training at both the undergraduate and graduate levels, having taken a variety of courses within their discipline. Like experts in other fields, drama experts have an extensive domain-specific knowledge base. In the course of their training they have studied the history of theater and have been exposed to a broad spectrum of dramatic literature, gaining an appreciation for the movements and trends that have been pivotal in the development of theater. They have learned the techniques of play analysis and production and have studied theories of acting and directing. Most importantly, they have mastered the technical vocabulary of drama,

which Perry (1984) characterizes as a highly specialized use of language involving "the personal articulation of things very difficult to say" (p. 30). In short, they have learned how to think and talk about drama.

The experts in this study also had extensive experience in their field. All but one expert had acting experience, many working in non-Equity and Equity productions. All the experts had casting or directing experience as well as experience teaching drama. Additionally, several had experience judging drama contests or working as drama critics. The experts often attended plays and movies, subscribed to a variety of theater magazines, and frequently purchased books about drama. How does acquiring experience in this field contribute to the development of expertise?

Through experience the drama expert builds a vast memory store of past performances. Smith (1970) characterizes the process as one of developing "an aesthetic conceptual net or map that facilitates storage and recall of relevant facts and procedures in the aesthetic domain" (p. 30). Perhaps drama experts' specialized knowledge is structured in hierarchically organized chunks, as has been shown in such diverse classes of experts as chess masters (Chase & Simon, 1973), expert physicists (Chi, Glaser & Rees, 1982), musicians (Sloboda, 1976), and electronic technicians (Egan & Schwartz, 1979). Chunking creates a rich knowledge of utility, efficiently and logically connected. Just as the chess master can store in memory over 50,000 images of chessboards that can be accessed to help decide which move to make in a game, the drama expert may store in memory details of past performances which could be accessed when judging an actor's ability.

In his study of 300 American theater critics Comtois (1977) found that the average critic had attended about 900 plays. Using Comtois' findings, one might estimate that the average critic had spent about 2,700 hours viewing

plays and considerably more time reading, thinking, and writing about the plays to produce reviews. It seems reasonable to assume that the drama teachers and casting directors in this study had similarly logged many hours watching performances and evaluating actors' abilities. By viewing a large number of performances, the expert develops a sense of what distinguishes a "good" performance from a "bad" one and acquires what Altschuler and Janaro (1967) term "a vision of the ideal." The expert comes to understand the standards or criteria that other experts use in making judgments in drama and adopts those standards which embody his/her own "vision of the ideal."

The expert has seen actors of various levels of ability and recognizes that not all of them are capable of attaining his "vision of the ideal," so the expert adjusts his/her expectations of each actor's capabilities taking into consideration factors such as the amount of training the actor has had, the actor's age, etc. Drama critic John Simon (as quoted in Searle, 1974) describes the importance of developing a "sliding scale" of excellence:

> A critical standard has to be both uniform and subdivisible. That is to say, in a sense you have a solid ideal of what you think is excellence. But, in another sense, you have a sliding scale and adjust it to the type of thing you are seeing . . . you sort of automatically evolve a sense of what might be the best that such a group could do, in your opinion, and then you judge according to that. (p. 11)

In Simon's view, the expert does not use different standards to judge persons of varying levels of ability. Rather, for each standard the expert can define low, medium, and high performance levels that reflect the expert's knowledge of what is the most and what is the least one can expect of an actor of a certain level of ability.

The drama expert has learned to identify the critical aspects of a performance. While the expert is unable to view all aspects simultaneously, he/she knows which ones to attend to. The expert can focus upon a number of

aspects in a single viewing and can take into consideration multiple criteria when assessing the quality of a performance. Because the expert knows what to look for, his/her judgments show strong intra-judge consistency from one occasion to the next.

Contributions to the literature

This study makes several unique contributions to the literature on the nature of expertise in aesthetic judgment. First, the study investigated expertise in making judgments about acting ability. The vast majority of prior research on this topic has been in the visual arts, not the performing arts. Thus, the study extends the scope of research to a different arts domain. Second, the study moves us beyond the narrow focus of past research on inter-judge agreement as a criterion for expertise. In this study other criteria were identified which proved useful in differentiating the ratings of experts from those of buffs and novices. Third, in the past there have been few published attempts to measure acting ability. No instruments for this purpose are commercially available. Drama teachers have had few tools and techniques to draw upon when faced with the difficult task of evaluating students' progress in the classroom. This study stands as a pioneer attempt to employ a model for objective measurement when designing an instrument to evaluate high school students' monologue performances. Fourth, past research has focused on comparing groups at both ends of the continuum--experts and novices. In the present study an intermediate group (i.e., theater buffs) was included which made it possible to study the transition from novice to expert. Finally, the present study employed data analysis techniques unlike those used in other studies of expertise in aesthetic judgment. Modeling the problem as a many-faceted situation provided the means for investigating each of the facets independent of the other facets, making objective measurement possible.

References

Altschuler, T., & Janaro, R. P. (1967). *Responses to drama: An introduction to plays and movies*. New York: Houghton Mifflin.

Bamossy, G., Johnston, M., & Parsons, M. (1985). The assessment of aesthetic judgment ability. *Empirical Studies of the Arts, 3*(1), 63-79.

Beard, A. D. (1978). The quantification of visual aesthetic judgment using verbal and nonverbal scales (Doctoral dissertation, University of Chicago). *Dissertation Abstracts International, 39*, 3789A.

Berlyne, D. E., & Ogilvie, J. C. (1974). Dimensions of perception of paintings. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics*. Washington, DC: Hemisphere.

Burt, C. (1934). The psychology of art. In *How the Mind Works* (pp. 267-310). New York: D. Appleton-Century Co.

Cattell, J., Glascock, J., & Washburn, M. F. (1918). Experiments on a possible test of aesthetic judgment of pictures. *American Journal of Psychology, 29*(3), 333-336.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of intelligence*. Hillsdale, NJ: Erlbaum.

Child, I. L. (1962). Personal preferences as an expression of aesthetic sensitivity. *Journal of Personality, 30*, 496-512.

Child, I. L. (1968). Esthetics. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology: Vol. 3. The Individual in a Social Context* (2nd ed.) (pp. 853-916). Reading, MA: Addison-Wesley Publishing Co.

Child, I. L. (1972). Esthetics. In P. H. Mussen & M. R. Rosenzweig (Eds.), *Annual Review of Psychology: Vol. 23* (pp. 669-694). Palo Alto, CA: Annual Reviews, Inc.

Comtois, M. E. (1977). Profile of the critic. In M. E. Comtois & L. F. Miller (Comps.), *Contemporary American theater critics: A directory and anthology of their works*. Metuchen, NJ: Scarecrow Press, Inc.

Dewar, H. (1938). A comparison of tests of artistic appreciation. *British Journal of Educational Psychology, 8*, 149-158.

Egan, D. E., & Schwartz, B. (1979). Chunking in recall of symbolic drawings. *Memory and Cognition, 7*, 149-158.

Einhorn, H. L (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59*(5), 562-571.

Einhorn, H. L., & Koelb, C. (1982). A psychometric study of literary-critical judgment. *Modern Language Studies, 12*(3), 59-82.

Eysenck, H. J. (1972). Personal preferences, aesthetic sensitivity and personality in trained and untrained subjects. *Journal of Personality, 40*(4), 544-557.

Farnsworth, P. R. (1969). Nature of musical taste. In *The social psychology of music* (2nd ed.) (pp. 97-133). Ames, IA: The Iowa State University Press.

Frances, R., & Voillaume, H. (1964). Une composante du jugement pictural: La fidelite de la representation. *Psychologie Francaise, 9*, 241-256.

Getzels, J. W., & Csikszentmihalyi, M. (1969). Aesthetic opinion: An empirical study. *Public Opinion Quarterly, 33*(1), 34-45.

Gordon, D. A. (1952). Methodology in the study of art evaluation. *Journal of Aesthetics and Art Criticism, 10*(4), 338-352.

Gordon, K. A. (1923). A study of esthetic judgments. *Journal of Experimental Psychology, 6*(1), 36-43.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hussain, F. (1966). L'unite des criteres de jugement esthetique: Enquete experimentale sur la correspondance des arts. *Psychologie Francaise, 11*(3), 285-298.

Landy, F. L., & Farr, J. L. (1983). The measurement of work performance: Methods, theory and applications. New York: Academic Press.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. (Doctoral dissertation, University of Chicago). *Dissertation Abstracts International, 50*, 2029A.

Myford, C. M. (1989). *The nature of expertise in aesthetic judgment: Beyond inter-judge agreement*. (Doctoral dissertation, University of Chicago). *Dissertation Abstracts International, 50*, 3562A.

O'Hare, D. (1976). Individual differences in perceived similarity and preference for visual art: A multidimensional scaling analysis. *Perception and Psychophysics, 20*(6), 445-452.

Perry, L. R. (1984). The arts judgment and language. *Journal of Aesthetic Education, 18*(1), 21-33.

Searle, J. (1974). Four drama critics. *Drama Review, 18*(3), 5-23.

Skager, R. W., Schultz, C. B., & Klein, S. P. (1966). Points of view about preference as tools in the analysis of creative products. *Perceptual and Motor Skills, 22*(1), 83-94.

Sloboda, J. A. (1976). Visual perception of musical notation: Registering pitch symbols in memory. *Quarterly Journal of Experimental Psychology, 28,* 1-16.

Smith, R. A. (1970). Psychology and aesthetic education. *Studies in Art Education, 11*(3), 20-30.

Valentine, C. W. (1962). *The experimental psychology of beauty.* London: Metheun and Co., Ltd.

Voillaume, H. (1965). Les activites picturales des enfants et les reactions comparees des enfants et des adultes devant les oeuvres d'enfants. *Psychologie Francaise,* 10, 178-187.

Wilson, B. (1970). The relationship between years of art training and the use of aesthetic judgmental criteria among high school students. *Studies in Art Education, 13*(2), 34-43.

Winner, E. (1982). *Invented worlds: The psychology of the arts.* Cambridge, MA: Harvard University Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

TABLE 1

PAIRWISE TESTS FOR RATING CONSISTENCY TO INVESTIGATE BETWEEN-GROUP DIFFERENCES IN CONTESTANT MEASURES

Groups	*Time 1* χ^2_s	*Time 2* χ^2_s
Expert vs. Buff	157.71*	104.96*
Expert vs. Novice	461.63*	531.81*
Buff vs. Novice	313.29*	285.05*

* $p < .005$

TABLE 2

DIFFERENCES BETWEEN CONTESTANT MEASURES
FOR EXPERTS, BUFFS, AND NOVICES--TIME 1

Contestant	Expert Calibration	Buff Calibration	Novice Calibration	Exp/Buff *z*	Buff/Nov *z*	Exp/Nov *z*
Mercutio	-0.80	-0.22	-0.59	-8.20**	5.23**	-2.97**
Ophelia	0.48	0.65	0.93	-2.40*	-3.59**	-5.76**
Mark Antony	0.58	0.68	1.78	-1.41	-11.66**	-12.72**
Juliet	0.04	0.18	-0.49	-1.98*	9.48**	7.50**
Lady Anne	-0.24	0.23	0.70	-6.65**	6.02**	-12.04**
Caliban	0.60	0.99	0.98	-4.99**	0.12	-4.87**
Iago	0.67	0.52	0.51	2.12*	0.13	2.05*
Paulina	-0.27	-0.43	-0.69	2.26*	3.33**	5.38**

*$p < .05$
**$p < .01$

TABLE 3

DIFFERENCES BETWEEN CONTESTANT MEASURES
FOR EXPERTS, BUFFS, AND NOVICES--TIME 2

Contestant	Expert Calibration	Buff Calibration	Novice Calibration	Exp/Buff *z*	Buff/Nov *z*	Exp/Nov *z*
Mercutio	-0.70	-0.42	-0.49	-3.96**	0.90	-2.69**
Ophelia	0.67	1.21	1.72	-6.28**	-5.15**	-12.21**
Mark Antony	0.64	0.60	1.61	0.57	-11.74**	-11.28**
Juliet	0.17	0.18	-0.03	-0.14	2.69**	2.56*
Lady Anne	-0.24	0.10	0.92	-4.81**	-10.50**	-14.65**
Caliban	0.78	1.12	1.07	-4.35**	0.59	-3.71**
Iago	0.49	0.67	0.62	-2.55*	0.64	-1.66
Paulina	-0.44	-0.51	-0.61	0.99	1.28	2.15*

*$p < .05$
**$p < .01$

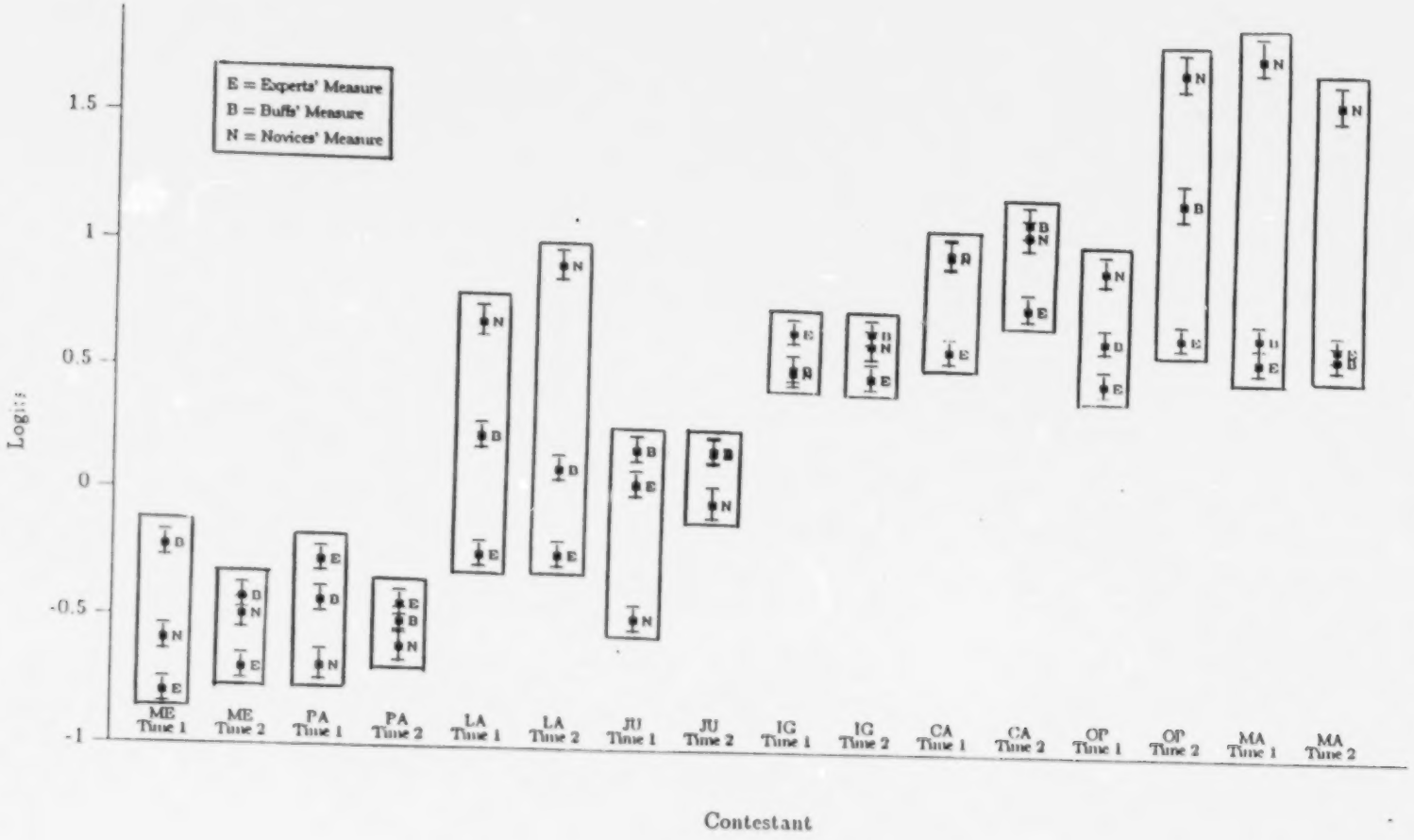

Figure 1. A comparison of novices', buffs', and experts' measures of contestant ability at Time 1 and Time 2.

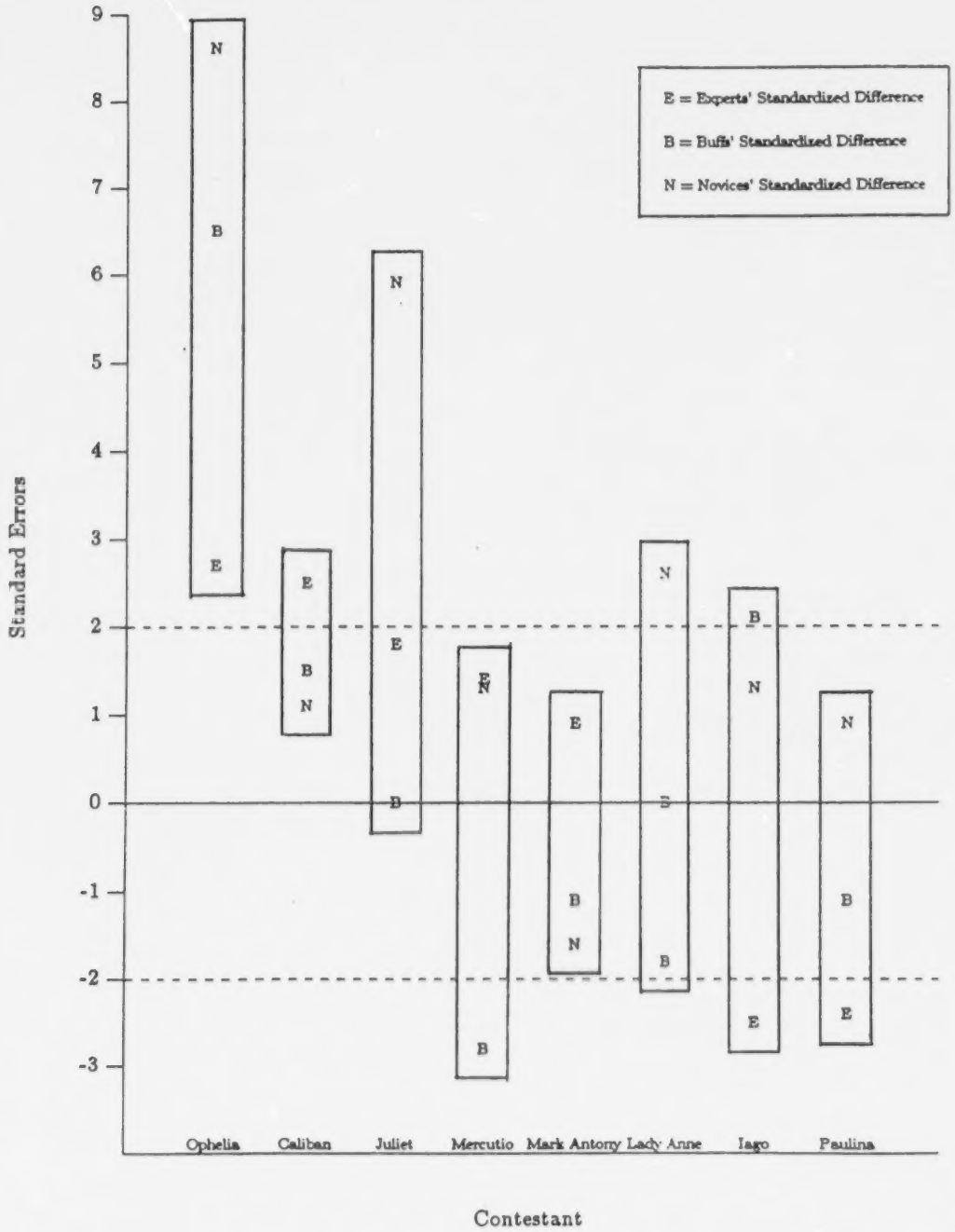

Figure 2. A comparison of changes in measures of contestant ability from Time 1 to Time 2.